Bacterial core gene set
Washington University Genome Center

Author: Makedonka Mitreva
Version: 1.01
Effective Date: 12/15/08

1 Abstract

None

2 Introduction

The bacterial core set is used in testing the completeness of draft genomes. The core set comprises of conserved single copy genes from *Escherichia coli* str. K-12 substr. MG1655, *Rhodobacter sphaeroides* 2.4.1, *Treponema pallidum* subsp. pallidum str. Nichols and *Staphylococcus aureus* subsp. Aureus. Coverage statistic is calculated by the percentage of the core set that have orthologs in the draft genome.

3 Requirements

3.1 Data requirements

Input fasta file: Input is a protein fasta format file or the assembly sequences.

Core genes: This is the core set of 200 genes. This is also a fasta file.

Core genes cluster file: Each line of the cluster file contains core genes that are orthologs. There are 66 groups and each group will contain at least 1 gene and at most 4 genes.

3.2 Software requirements

perl : Any Perl installation

blast or other alignment program : NCBI BLAST or WU- BLAST or other alignment software

get_coregroups_coverage.pl : Perl script to calculate coverage of core genes

3.2 Compute requirements

None

Author: Makedonka Mitreva
Version: 1.01
Effective Date: 12/15/08

4 Procedure

1. Alignments to core genes

Using the core genes file as query and the input fasta file as subject, obtain alignments to the 200 core genes. Any alignment software can be used.

(Example: blastx with parameters " hitdist=40 wordmask=seg postsw topcomboN=1")

2. Filtering alignments

All alignments that have at least than 30% identity over at least 30% of the length of the core genes are considered valid. The core genes that have alignments meeting these criteria are collected to be used as input for the next step.

3. Coverage of core groups

The list of core genes that had homology to the input fasta sequences with 30% identity over 30% length are passed on to this script – get_coregroups_coverage.pl

Script run with the following parameters :

get_coregroups_coverage.pl

-coregroups <cluster file>

-gene_list <file with core genes with valid alignments>

The gene list is the list of core genes from step 2.

The coregroups parameters is the core group cluster file.

5 Implementation

The script takes the orthologs for the core genes in the draft genomes as input and searches against the core groups to identify the number of groups that have at least one of the genes. The percentage of core groups identified is printed as output.

6 Discussion

Genes from 4 finished genomes - Escherichia coli str. K-12 substr. MG1655 (4,131 genes), Rhodobacter sphaeroides 2.4.1 (4,242 genes), Treponema pallidum subsp. pallidum str. Nichols (1,031 genes) and Staphylococcus aureus subsp. aureus USA300(2,683 genes) were clustered using OrthoMCL[1]. These 12,087 genes from all 4 species grouped into putative ortholog groups, first by searching all the sequences against each other using BLAST and applying a markov cluster algorithm on the results. Clustering with the default parameters (Inflation factor 1.5) identified 290 groups with orthologous genes in all 4 organisms. Of these 235 groups were with single copy genes per genome. An earlier study by Callister et al. [2] with 17 environmental and pathogenic bacterial genomes identified 229 core genes out of which 111 were single copy genes. Ninety-nine of these genes were identifiable in the 4 test genomes using BLAST criterion used in assembly assessment and analysis.

The common core set with 99 genes was refined by examining Gene Ontology(GO) annotations, verifying coverage with 621 finished bacterial genomes and validation with 34 Human Gut Microbiome Initiative (HGMI) genomes. This refinement reduced the core set to 66 genes and all of the 66 are present in the 621 finished genomes. Functional categories enriched in the core gene set are protein synthesis, biosynthesis of prosthetic groups and carriers, nucleotide biosynthesis, amino acid biosynthesis, protein transport etc.

To generate non-redundant list of genes (minimum number of genes per group that covers maximum number of genomes), we started with the 264 (66x 4) genes from the 66 groups, and removed genes that did not give any additional information. A total of 64 genes were removed and 200 core genes remained from the 66 core

groups. Coverage of the draft genomes can be estimated by identifying which of the 66 groups have orthologs.

Evaluating if genes from these 66 core groups are present in the draft genomes gives a good indication of the completeness and coverage of the draft genomes.

A more detailed description of Archaeal core gene selection and evaluation can be found in Abubucker & Mitreva Bacterial core gene evaluation.

7 Related Documents & References

[1] Li Li, Christian J Stoeckert, Jr. and Davis S. Roos (2003), OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes, *Genome Res. 2003. 13: 2178-2189*

[2] Callister SJ, McCue LA, Turse JE, Monroe ME, Auberry KJ, et al. (2008) Comparative Bacterial Proteomics: Analysis of the Core Genome Concept. PLoS ONE 3(2): e1542. doi:10.1371/journal.pone.0001542

8 Revision History

This is an HMP_specific requirement, not included in the SIGS submission. Please be sure to update this when any changes as made, to help the DACC organize SOPs.

Version	Author/Reviewer	Date	Change Made
1.01	Makedonka Mitreva	12/15/08	Establish SOP